In [41]:
```
//Perform the following operations using Python on the Air quality and H
a. Data cleaning
b. Data integration
c. Data transformation
d. Error correcting
e. Data model building
```

```
  Cell In [41], line 1
    //Perform the following operations using Python on the Air quality
and Heart Diseases data sets
    ^
SyntaxError: invalid syntax
```

In [42]:
```
import pandas as pd
import numpy as np
df = pd.read_csv("city_day.csv")
df
```

Out[42]:

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Ahmedabad | 2015-01-01 | NaN | NaN | NaN | 18.22 | 17.15 | NaN | 0.92 | 27.64 | 13 |
| **1** | Ahmedabad | 2015-01-02 | NaN | NaN | NaN | 15.69 | 16.46 | NaN | 0.97 | 24.55 | 3 |
| **2** | Ahmedabad | 2015-01-03 | NaN | NaN | NaN | 19.30 | 29.70 | NaN | 17.40 | 29.07 | 3 |
| **3** | Ahmedabad | 2015-01-04 | NaN | NaN | NaN | 18.48 | 17.97 | NaN | 1.70 | 18.59 | 3 |
| **4** | Ahmedabad | 2015-01-05 | NaN | NaN | NaN | 21.42 | 37.76 | NaN | 22.10 | 39.33 | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **25550** | Thiruvananthapuram | 2020-04-06 | 16.88 | 31.28 | 4.32 | 10.18 | 13.71 | 5.55 | 0.32 | 4.24 | 3 |
| **25551** | Thiruvananthapuram | 2020-04-07 | 18.57 | 35.34 | 5.70 | 9.56 | 12.70 | 4.97 | 0.46 | 6.00 | 3 |
| **25552** | Thiruvananthapuram | 2020-04-08 | 18.85 | 37.62 | 5.44 | 8.86 | 12.09 | 4.39 | 0.47 | 5.79 | 5 |
| **25553** | Thiruvananthapuram | 2020-04-09 | 19.06 | 45.18 | 3.37 | 7.91 | 9.92 | 4.37 | 0.52 | 5.76 | 4 |
| **25554** | Thiruvananthapuram | 2020-04-10 | 23.44 | 42.10 | 2.84 | 7.18 | 8.93 | 3.49 | 0.52 | 5.57 | 3 |

25555 rows × 16 columns

In [43]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25555 entries, 0 to 25554
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   City        25555 non-null  object 
 1   Date        25555 non-null  object 
 2   PM2.5       21300 non-null  float64
 3   PM10        14853 non-null  float64
 4   NO          21046 non-null  float64
 5   NO2         22396 non-null  float64
 6   NOx         21534 non-null  float64
 7   NH3         15770 non-null  float64
 8   CO          23556 non-null  float64
 9   SO2         22056 non-null  float64
 10  O3          22045 non-null  float64
 11  Benzene     20433 non-null  float64
 12  Toluene     18146 non-null  float64
 13  Xylene      9138 non-null   float64
 14  AQI         21304 non-null  float64
 15  AQI_Bucket  21304 non-null  object 
dtypes: float64(13), object(3)
memory usage: 3.1+ MB
```

In [44]:
```
df.isnull().sum()
```

Out[44]:
```
City             0
Date             0
PM2.5         4255
PM10         10702
NO            4509
NO2           3159
NOx           4021
NH3           9785
CO            1999
SO2           3499
O3            3510
Benzene       5122
Toluene       7409
Xylene       16417
AQI           4251
AQI_Bucket    4251
dtype: int64
```

In [45]:
```
df.head(6)
```

Out[45]:

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 | O3 | Benzer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Ahmedabad | 2015-01-01 | NaN | NaN | NaN | 18.22 | 17.15 | NaN | 0.92 | 27.64 | 133.36 | 0.( |
| **1** | Ahmedabad | 2015-01-02 | NaN | NaN | NaN | 15.69 | 16.46 | NaN | 0.97 | 24.55 | 34.06 | 3.( |
| **2** | Ahmedabad | 2015-01-03 | NaN | NaN | NaN | 19.30 | 29.70 | NaN | 17.40 | 29.07 | 30.70 | 6.8 |
| **3** | Ahmedabad | 2015-01-04 | NaN | NaN | NaN | 18.48 | 17.97 | NaN | 1.70 | 18.59 | 36.08 | 4.4 |
| **4** | Ahmedabad | 2015-01-05 | NaN | NaN | NaN | 21.42 | 37.76 | NaN | 22.10 | 39.33 | 39.31 | 7.( |
| **5** | Ahmedabad | 2015-01-06 | NaN | NaN | NaN | 38.48 | 81.50 | NaN | 45.41 | 45.76 | 46.51 | 5.4 |

In [46]:
```
df.columns
```

Out[46]:
```
Index(['City', 'Date', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'C
O', 'SO2',
       'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI', 'AQI_Bucket'],
      dtype='object')
```

In [47]:
```
data2=df.copy()
```

In [48]:
```
data2=data2.fillna(data2.mean())      #replace null values with mean
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_13716\2591363272.py:1: Fut
ureWarning:

The default value of numeric_only in DataFrame.mean is deprecated. In
a future version, it will default to False. In addition, specifying 'n
umeric_only=None' is deprecated. Select only valid columns or specify
the value of numeric_only to silence this warning.
```

In [49]:
```
data2.head()
```

Out[49]:

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | S |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Ahmedabad | 2015-01-01 | 72.824121 | 127.653183 | 17.658008 | 18.22 | 17.15 | 25.66772 | 0.92 | 27 |
| **1** | Ahmedabad | 2015-01-02 | 72.824121 | 127.653183 | 17.658008 | 15.69 | 16.46 | 25.66772 | 0.97 | 24 |
| **2** | Ahmedabad | 2015-01-03 | 72.824121 | 127.653183 | 17.658008 | 19.30 | 29.70 | 25.66772 | 17.40 | 29 |
| **3** | Ahmedabad | 2015-01-04 | 72.824121 | 127.653183 | 17.658008 | 18.48 | 17.97 | 25.66772 | 1.70 | 18 |
| **4** | Ahmedabad | 2015-01-05 | 72.824121 | 127.653183 | 17.658008 | 21.42 | 37.76 | 25.66772 | 22.10 | 39 |

In [50]:
```
data2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25555 entries, 0 to 25554
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   City        25555 non-null  object 
 1   Date        25555 non-null  object 
 2   PM2.5       25555 non-null  float64
 3   PM10        25555 non-null  float64
 4   NO          25555 non-null  float64
 5   NO2         25555 non-null  float64
 6   NOx         25555 non-null  float64
 7   NH3         25555 non-null  float64
 8   CO          25555 non-null  float64
 9   SO2         25555 non-null  float64
 10  O3          25555 non-null  float64
 11  Benzene     25555 non-null  float64
 12  Toluene     25555 non-null  float64
 13  Xylene      25555 non-null  float64
 14  AQI         25555 non-null  float64
 15  AQI_Bucket  21304 non-null  object 
dtypes: float64(13), object(3)
memory usage: 3.1+ MB
```

In [51]:
```
dist=(data2['City'])
distset=set(dist)
dd=list(distset)
dictOfWords = { dd[i] : i for i in range(0, len(dd) )}
data2['City']=data2['City'].map(dictOfWords)
```

In [52]:
```
dist=(data2['AQI_Bucket'])
distset=set(dist)
dd=list(distset)
dictOfWords = { dd[i] : i for i in range(0, len(dd) )}
data2['AQI_Bucket']=data2['AQI_Bucket'].map(dictOfWords)
```

In [53]:
```
data2["AQI_Bucket"]=data2["AQI_Bucket"].fillna(data2["AQI_Bucket"].mean(
```

In [54]:
```
data2
```

Out[54]:

| | City | Date | PM2.5 | PM10 | NO | NO2 | NOx | NH3 | CO | SO2 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 13 | 2015-01-01 | 72.824121 | 127.653183 | 17.658008 | 18.22 | 17.15 | 25.66772 | 0.92 | 27.64 |
| **1** | 13 | 2015-01-02 | 72.824121 | 127.653183 | 17.658008 | 15.69 | 16.46 | 25.66772 | 0.97 | 24.55 |
| **2** | 13 | 2015-01-03 | 72.824121 | 127.653183 | 17.658008 | 19.30 | 29.70 | 25.66772 | 17.40 | 29.07 |
| **3** | 13 | 2015-01-04 | 72.824121 | 127.653183 | 17.658008 | 18.48 | 17.97 | 25.66772 | 1.70 | 18.59 |
| **4** | 13 | 2015-01-05 | 72.824121 | 127.653183 | 17.658008 | 21.42 | 37.76 | 25.66772 | 22.10 | 39.33 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **25550** | 17 | 2020-04-06 | 16.880000 | 31.280000 | 4.320000 | 10.18 | 13.71 | 5.55000 | 0.32 | 4.24 |
| **25551** | 17 | 2020-04-07 | 18.570000 | 35.340000 | 5.700000 | 9.56 | 12.70 | 4.97000 | 0.46 | 6.00 |
| **25552** | 17 | 2020-04-08 | 18.850000 | 37.620000 | 5.440000 | 8.86 | 12.09 | 4.39000 | 0.47 | 5.79 |
| **25553** | 17 | 2020-04-09 | 19.060000 | 45.180000 | 3.370000 | 7.91 | 9.92 | 4.37000 | 0.52 | 5.76 |
| **25554** | 17 | 2020-04-10 | 23.440000 | 42.100000 | 2.840000 | 7.18 | 8.93 | 3.49000 | 0.52 | 5.57 |

25555 rows × 16 columns

In [55]:
```
data2.isnull().sum()
```

Out[55]:
```
City          0
Date          0
PM2.5         0
PM10          0
NO            0
NO2           0
NOx           0
NH3           0
CO            0
SO2           0
O3            0
Benzene       0
Toluene       0
Xylene        0
AQI           0
AQI_Bucket    0
dtype: int64
```

In [56]:
```
data2 = data2.drop('Date' , 1)
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_13716\114922541.py:1: Futu
reWarning:

In a future version of pandas all arguments of DataFrame.drop except f
or the argument 'labels' will be keyword-only.
```

In [57]:
```
data2.columns
```

Out[57]:
```
Index(['City', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO
2', 'O3',
       'Benzene', 'Toluene', 'Xylene', 'AQI', 'AQI_Bucket'],
      dtype='object')
```

In [58]:
```
data2 =data2.drop('AQI_Bucket', 1)
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_13716\1146162600.py:1: Fut
ureWarning:

In a future version of pandas all arguments of DataFrame.drop except f
or the argument 'labels' will be keyword-only.
```

In [59]:
```
import plotly.express as px
#EDA (Analyse the data)
fig = px.scatter(df, x="City", y="AQI")    #Plotting the Bubble Chart
fig.show()
```

In [60]:
```
import plotly
```

In [61]:
```
pip install plotly
```

```
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: plotly in c:\users\admin\appdata\local
\programs\python\python310\lib\site-packages (5.14.1)
Requirement already satisfied: tenacity>=6.2.0 in c:\users\admin\appda
ta\local\programs\python\python310\lib\site-packages (from plotly) (8.
2.2)
Requirement already satisfied: packaging in c:\users\admin\appdata\loc
al\programs\python\python310\lib\site-packages (from plotly) (21.3)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in c:\users\ad
min\appdata\local\programs\python\python310\lib\site-packages (from pa
ckaging->plotly) (3.0.9)


[notice] A new release of pip available: 22.2.2 -> 23.1.2
[notice] To update, run: C:\Users\Admin\AppData\Local\Programs\Python
\Python310\python.exe -m pip install --upgrade pip
```

In [62]:
```
import plotly.express as px
#EDA (Analyse the data)
fig2 = px.scatter(df,x="PM10",y="AQI")    #Plotting the Bubble Chart
fig2.show()
```

In [63]:
```
import plotly.express as px
#EDA (Analyse the data)
fig2 = px.scatter(df,x="NOx",y="AQI")    #Plotting the Bubble Chart
fig2.show()
```

In [64]:
```
data2.columns
```

Out[64]:
```
Index(['City', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO
2', 'O3',
       'Benzene', 'Toluene', 'Xylene', 'AQI'],
      dtype='object')
```

In [65]:
```
import plotly.express as px
#EDA (Analyse the data)
fig2 = px.scatter(df,x="CO",y="AQI")    #Plotting the Bubble Chart
fig2.show()
```

In [66]:
```
features = data2[['City', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'C
       'Benzene', 'Toluene', 'Xylene']]
labels = data2['AQI']
```

In [67]:
```
import plotly.express as px
#EDA (Analyse the data)
fig2 = px.scatter(df,x="Xylene",y="AQI")    #Plotting the Bubble Chart
fig2.show()
```

In [68]:
```
#splitting into train & test data
from sklearn.model_selection import train_test_split
Xtrain, Xtest, Ytrain, Ytest = train_test_split(features,labels,test_siz
```

In [69]:
```
pip install scikit-learn
```

```
Requirement already satisfied: scikit-learn in c:\users\admin\appdata
\local\programs\python\python310\lib\site-packages (1.1.3)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\admin
\appdata\local\programs\python\python310\lib\site-packages (from sciki
t-learn) (3.1.0)
Requirement already satisfied: numpy>=1.17.3 in c:\users\admin\appdata
\local\programs\python\python310\lib\site-packages (from scikit-learn)
(1.23.4)
Requirement already satisfied: scipy>=1.3.2 in c:\users\admin\appdata
\local\programs\python\python310\lib\site-packages (from scikit-learn)
(1.9.3)
Requirement already satisfied: joblib>=1.0.0 in c:\users\admin\appdata
\local\programs\python\python310\lib\site-packages (from scikit-learn)
(1.2.0)
Note: you may need to restart the kernel to use updated packages.


[notice] A new release of pip available: 22.2.2 -> 23.1.2
[notice] To update, run: C:\Users\Admin\AppData\Local\Programs\Python
\Python310\python.exe -m pip install --upgrade pip
```

In [70]:
```
#splitting into train & test data
from sklearn.model_selection import train_test_split
Xtrain, Xtest, Ytrain, Ytest = train_test_split(features,labels,test_siz
```

In [71]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import classification_report
from sklearn import metrics
from sklearn import tree
```

In [72]:
```
pip install matplotlib
```

```
Requirement already satisfied: matplotlib in c:\users\admin\appdata\lo
cal\programs\python\python310\lib\site-packages (3.6.2)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\admin\app
data\local\programs\python\python310\lib\site-packages (from matplotli
b) (1.4.4)
Requirement already satisfied: pillow>=6.2.0 in c:\users\admin\appdata
\local\programs\python\python310\lib\site-packages (from matplotlib)
(9.3.0)
Requirement already satisfied: packaging>=20.0 in c:\users\admin\appda
ta\local\programs\python\python310\lib\site-packages (from matplotlib)
(21.3)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\admin\appd
ata\local\programs\python\python310\lib\site-packages (from matplotli
b) (1.0.6)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\admin\app
data\local\programs\python\python310\lib\site-packages (from matplotli
b) (4.38.0)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\admin
\appdata\local\programs\python\python310\lib\site-packages (from matpl
otlib) (2.8.2)
Requirement already satisfied: cycler>=0.10 in c:\users\admin\appdata
\local\programs\python\python310\lib\site-packages (from matplotlib)
(0.11.0)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\admin\appd
ata\local\programs\python\python310\lib\site-packages (from matplotli
b) (3.0.9)
Requirement already satisfied: numpy>=1.19 in c:\users\admin\appdata\l
ocal\programs\python\python310\lib\site-packages (from matplotlib) (1.
23.4)
Requirement already satisfied: six>=1.5 in c:\users\admin\appdata\loca
l\programs\python\python310\lib\site-packages (from python-dateutil>=
2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.


[notice] A new release of pip available: 22.2.2 -> 23.1.2
[notice] To update, run: C:\Users\Admin\AppData\Local\Programs\Python
\Python310\python.exe -m pip install --upgrade pip
```

In [73]:
```
from sklearn.ensemble import RandomForestRegressor
from sklearn.datasets import make_regression
regr = RandomForestRegressor(max_depth=2, random_state=0)
regr.fit(Xtrain,Ytrain)
print(regr.predict(Xtest))
```

```
[126.53394437 126.53394437 128.27373545 ... 126.53394437 126.53394437
 126.53394437]
```

In [74]:
```
y_pred=regr.predict(Xtest)
```

In [75]:
```
from sklearn.metrics import r2_score
```

r2_score(Ytest,y_pred)

In [76]:
```
r2_score(Ytest, y_pred)
```

Out[76]: 0.6627161490108119

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: